JAMIA Open, 2024, **7(2)**, ooae041
https://doi.org/10.1093/jamiaopen/ooae041
Research and Applications

Research and Applications

Validation and clinical discovery demonstration of breast cancer data from a real-world data extraction platform

Amanda Nottke, PhD*[,1], Sophia Alan, RN, BSN[1], Elise Brimble, MSc, MS, CGC[1], Anthony B. Cardillo, MD[1], Lura Henderson, MSN, APRN[1], Hana E. Littleford, PhD[1], Susan Rojahn, PhD[2], Heather Sage, RN[1], Jessica Taylor, MSN, CNS[1], Lisandra West-Odell, PhD[1], Alexandra Berk, MA[1]

[1]Ciitizen, San Francisco, CA 94112, United States, [2]Invitae, San Francisco, CA 94103, United States

*Corresponding author: Amanda Nottke, PhD, Ciitizen, 108 Justin Dr, San Francisco, CA 94112, United States (amanda.nottke@ciitizen.com)

Abstract

Objective: To validate and demonstrate the clinical discovery utility of a novel patient-mediated, medical record collection and data extraction platform developed to improve access and utilization of real-world clinical data.

Materials and Methods: Clinical variables were extracted from the medical records of 1011 consented patients with breast cancer. To validate the extracted data, case report forms completed using the structured data output of the platform were compared to manual chart review for 50 randomly-selected patients with metastatic breast cancer. To demonstrate the platform's clinical discovery utility, we identified 194 patients with early-stage clinical data who went on to develop distant metastases and utilized the platform-extracted data to assess associations between time to distant metastasis (TDM) and early-stage tumor histology, molecular type, and germline *BRCA* status.

Results: The platform-extracted data for the validation cohort had 97.6% precision (91.98%-100% by variable type) and 81.48% recall (58.15%-95.00% by variable type) compared to manual chart review. In our discovery cohort, the shortest TDM was significantly associated with metaplastic (739.0 days) and inflammatory histologies (1005.8 days), HR – /HER2 – molecular types (1187.4 days), and positive *BRCA* status (1042.5 days) as compared to other histologies, molecular types, and negative *BRCA* status, respectively. Multivariable analyses did not produce statistically significant results.

Discussion: The precision and recall of platform-extracted clinical data are reported, although specificity could not be assessed. The data can generate clinically-relevant insights.

Conclusion: The structured real-world data produced by a novel patient-mediated, medical record-extraction platform are reliable and can power clinical discovery.

Lay Summary

"Real-world data" (RWD) refers to information about health and health care as it occurs naturally over time and across treatment settings. Because RWD is not limited to specific individuals, locations, or treatments, RWD is particularly valuable for exploring how a disease occurs, progresses, and is treated in rarer diseases like cancers and inherited syndromes. However, the quality of RWD depends on how it is generated, and so it is critical to measure the accuracy and comprehensiveness of the data. The Ciitizen platform (hereinafter "the Platform") engages individual patients to generate de-identified RWD from medical records for research purposes. In this study, we show that these data are both precise and comprehensive as compared to the "gold standard" of manually collecting information from medical record documents. We also explore how these data can be used for research purposes by asking a specific question of interest for patients with breast cancer: how might clinical characteristics at the time of early stage breast cancer diagnosis be related to the time to develop more advanced (metastatic) disease. Taken together, this study supports both data quality validation and research utility of this novel patient-centered RWD platform.

Key words: real-world data; electronic health records; breast cancer; metastasis; validation studies; data accuracy; real-world evidence.

Introduction

Clinical real-world data (RWD) are observational data describing patient health and health care delivery. RWD can be collected from a variety of sources, including electronic health records (EHRs), insurance claims databases, disease registries, and patient surveys, and can be a valuable source of real-world evidence (RWE) for studies on patient outcomes, investigational therapies, and clinical best practices.[1–4] Because RWE reflects clinical decision-making not constrained by a study protocol, it may better capture the true experience of patients living with a particular disease or receiving a particular therapeutic intervention, as compared to data collected through clinical trials or other restrictive clinical research settings. Further, RWE can be assembled from cohorts with greater diversity in patient backgrounds and geography, which can be especially useful for studies involving patients with rare disease. Given the potential utility of RWE in drug development, the US Food and Drug Administration (FDA) has formally established guidance for using RWD in premarket and postmarket evaluations of

Received: June 6, 2023; **Revised:** February 29, 2024; **Editorial Decision:** April 26, 2024

novel treatments, including clinical studies to support regulatory approval decisions of new drugs or biological products, or new indications for existing ones.[5,6]

While both insurance claims and EHR data can provide patient health and care delivery data including diagnoses, prescriptions, and procedures, EHRs offer a richer and more nuanced source of RWD.[7] Radiographic images, genetic and molecular test results, pathology reports, and especially free text notes which may be part of the full medical record can contain highly valuable clinical insights not available in claims data. However, there are also several challenges with harnessing EHR data as a source of clinical evidence. Many components of EHRs, including the free-text notes from clinicians, are not standardized or structured, which can complicate automated extraction. Without automation, data extraction from EHRs requires time- and skill-intensive manual chart review by clinical experts and cannot be scaled to large cohorts. In addition, there are many EHR systems in use across healthcare entities in the United States, which adds complexity to data management and utilization. Given that many patients switch health care systems or insurance providers frequently over time, an EHR-based dataset may have missing data if taken from a single healthcare system. Finally, almost all secondary data sources contain de-identified patient data, which disrupts longitudinal tracking across institutions and linking of patient-reported outcomes to disease course and treatment data.

To address some of these gaps, we set out to build a novel patient-facing and patient-mediated RWD extraction platform that processes data from EHRs into structured and queryable data elements. Upon patient consent, their medical records (digital or paper) were collected and then analyzed by natural language processing algorithms to create structured data whose accuracy was confirmed by clinical experts via 2 rounds of human review. Centering this platform around patient consent had several benefits. Firstly, it enabled collection of their medical records across multiple institutions, thus minimizing any gaps in their medical history data. In addition, participants could be engaged prospectively over time to gather additional records to enrich the dataset. Participants also controlled access to both their structured data and their complete collection of medical records, which they could use as they saw fit, eg, for second opinions or clinical trial matching, and could be invited for specific research opportunities. Collectively, the processed data could be used for retrospective, observational research including regulated studies on novel therapies. RWD from this platform (hereinafter "the Platform") has already been used as natural history data to support a successful Investigational New Drug application for a novel treatment for pediatric patients with a severe form of epileptic encephalopathy.[8]

The objective of the study presented here is to use the classic measures of precision and recall to validate the accuracy and comprehensiveness of the Platform's performance as compared to manual chart review, the gold standard method for extracting data from medical records. A second objective is to demonstrate the potential for this data to generate clinical insights that may support hypotheses through a study of the time-to-distant metastasis (TDM) associated with specific tumor and patient characteristics. Both analyses leveraged datasets from patients with metastatic breast cancer, a disease with complex and evolving standards of care and an associated opportunity for RWE to contribute to the overall body of knowledge.

Methods

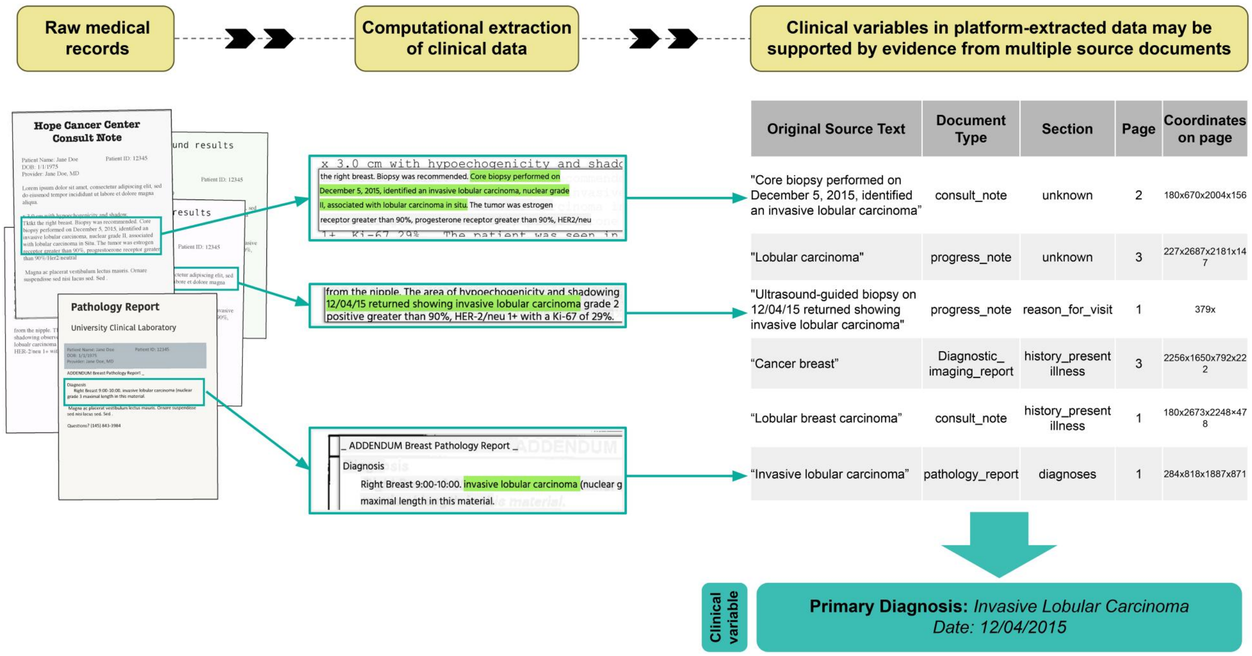
Overview of data extraction process

An overview of the data extraction process is shown in Figure 1. Following a directive from a consenting participant, the Health Insurance Portability and Accountability Act of 1996 (HIPAA) right of access is leveraged to obtain medical records from every institution reported to have been visited by the individual. A triage process guarantees recency of data and a minimum completeness of record collection. All medical records are stored electronically, and the participant can freely access them by logging into a secure portal. Medical record documents are then processed through natural language processing pipelines to convert unstructured data such as clinician notes and pathology reports into structured data consisting of clinical variables and their associated date(s) or date ranges that can be supported by multiple source documents (Figure 1). For example, a diagnosis of breast cancer on a particular date would be a single clinical variable that is supported by one or more medical records from multiple institutions. The automated extraction process includes document classification (eg, pathology report, progress report, genetic testing results) and structuring of clinical terms to a pre-established dictionary of data variable types specific to solid tumors (Table S1). In addition, modeling relationships connect certain variables, for example Adverse Events are modeled to the causative Medication, and Secondary Diagnoses (eg, metastasis to bone) are modeled to the Primary Diagnosis (eg, breast cancer). Following computational extraction, the structured data is reviewed by clinical experts who confirm clinical variables are accurate and generated modeling relationships where relevant, as supported by at least one source document for the term, associated date(s), and any causative association between variables. The resulting data elements are stored securely in a HIPAA-compliant, controlled access, indexed database.

Validation study

Validation study design

This validation study protocol received a determination of exempt status by Pearl IRB. Fifty patients with metastatic breast cancer who had provided research consent to using their de-identified medical records data were randomly selected from our processed metastatic breast cancer cases to be the validation cohort. Demographic and clinical characteristics were assessed using the platform data. Data extracted from medical records of these 50 selected patients was compared to chart review conducted by oncology nurse annotators (Figure S1), as in.[9] The oncology nurse annotators were contracted specifically for this analysis, in order to eliminate any bias due to familiarity with the extraction process or data structuring, and had comparable years of experience in clinical oncology and annotation. For 50 patients with metastatic breast cancer, 2 different annotators completed electronic case report forms (CRFs) using either raw medical record documents ("records direct") or the structured data produced by the platform ("platform CRF"). The platform CRF data was generated using the same set of raw medical records as the annotators used for the records direct condition. Variables on the CRF included primary breast cancer diagnosis,


Raw medical records
Computational extraction of clinical data
Clinical variables in platform-extracted data may be supported by evidence from multiple source documents
Hope Cancer Center Consult Note
Pathology Report
University Clinical Laboratory
Clinical variable
Primary Diagnosis: Invasive Lobular Carcinoma
Date: 12/04/2015

Original Source Text	Document Type	Section	Page	Coordinates on page
"Core biopsy performed on December 5, 2015, identified an invasive lobular carcinoma"	consult_note	unknown	2	180x670x2004x156
"Lobular carcinoma"	progress_note	unknown	3	227x2687x2181x147
"Ultrasound-guided biopsy on 12/04/15 returned showing invasive lobular carcinoma"	progress_note	reason_for_visit	1	379x
"Cancer breast"	Diagnostic_ imaging_report	history_present illness	3	2256x1650x792x222
"Lobular breast carcinoma"	consult_note	history_present illness	1	180x2673x2248×478
"Invasive lobular carcinoma"	pathology_report	diagnoses	1	284x818x1887x871

Figure 1. Platform overview. Clinical data is computationally extracted from raw medical records and confirmed by human review based on one or more supporting documents.

tumor histologic type, molecular type, stage, medications, disease statuses, and adverse events (Table S1). Dates were required for most variables on the CRF, with the exception of comorbidities which were allowed to be undated.

Validation study analysis

Completed platform CRFs for the same patient were compared with the records-direct CRFs serving as the reference standard. The comparison was completed in 2 phases. In the first phase, variables identified in both the records-direct CRF and the platform CRF were designated as true positives (TPs), variables identified in the records-direct CRF only were considered false negatives (FNs), and variables identified in the platform CRF only were considered false positives (FPs). However, as manual chart review is not necessarily error-proof, all FP and FN identified in the first phase were escalated to review by a third annotator. The escalation annotator reviewed source documents to determine if any variables in a platform CRF that were initially scored as FP could be verified in the source documents and were therefore missed in manual chart review. If the variables were verified in the source documentation, the score was adjusted from FP to TP. For any variables scored FN because they were identified in the records-direct CRF but not the platform CRF, a third annotator determined if any were out of scope for the study, such as diagnostic procedures done before the primary diagnosis, which the platform does not extract. If so, the score was considered to be manual chart review error and adjusted from FN to null (not scored). These adjustments enabled a more accurate comparison between the records-direct and platform conditions.

Two metrics were calculated to assess the accuracy and comprehensiveness of the extracted data: precision, calculated as the number of TP divided by the sum of TP and FP; and recall, calculated as the number of TP divided by the sum of TP and FN.

Demonstration of clinical utility

To assess the potential for the platform-extracted data to support clinical insights, we looked for correlations between clinical features and TDM among a cohort of patients with breast cancer from the platform database with distant metastasis (hereafter, the discovery cohort). From a starting population of 1011 research-consented patients with breast cancer, we identified those with medical records data at the time of primary breast cancer diagnosis and at least one of: germline *BRCA* status, HER2/HR molecular type, and histologic type, and with a latter date of documented metastasis to brain, bone, liver, and/or lung ($n = 194$). Separately, we examined the distribution of histologic types, molecular types, and stage at diagnosis among patients with invasive breast cancer in the Surveillance, Epidemiology, and End Results Program (SEER) cancer statistics registry (data from 1975-2017) to provide context for our platform-derived cohorts of 1011 as well as the 194 individuals in the discovery cohort.

Demographics and clinical characteristics were assessed in the full study cohort ($n = 1011$) as well as the discovery cohort ($n = 194$). Clinical characteristics included histologic type (ductal, lobular, ductal + lobular, inflammatory, and metaplastic) and molecular type (based on combinations of hormone receptor positive [HR+], hormone receptor negative [HR−], human epidermal growth factor receptor 2 positive [HER2+], and human epidermal growth factor receptor 2 negative [HER2−]). The time to distant metastasis (TDM) was defined as the time between the first primary diagnosis and the first metastatic diagnosis at each site, for example, metastasis to lung. For patients who had metastases to multiple sites, the first instance of metastasis to each site was calculated separately, in order to enable analyses based on site of disease. The average TDM was calculated across the discovery cohort and stratified by histologic type, molecular type, germline *BRCA* status, and site of distant metastasis.

Bivariate differences in average TDM were determined by unpaired, 2-tailed *t*-test with Welch's correction. Multivariable analysis was performed using ANOVA and Tukey's Honestly-Significant-Difference. Statistically significant findings were those with a *P*-value <.05.

Results

Validation study: agreement between data extraction platform and manual chart review

Among the 50 patients with metastatic breast cancer in the validation cohort, the mean age at diagnosis was 47.7 (SD: 9.3; Range: 29-69), the most common histology was Invasive ductal carcinoma (62%), and the most common molecular type was HR + /HER2 − (68%) (Table 1). Initial comparisons of CRFs completed with platform data and those completed directly from the medical records revealed 4089 variables for comparison (Table S2). Escalated review to validate or correct all FNs and FPs by a third annotator revealed 798 variables that were either putative FPs that were confirmed in the platform data source documents and missed in the manual chart review (and therefore re-coded as TPs), or putative FNs that were outside of the scope of the study (eg, a confirmed benign lump in the other breast prior to the primary diagnosis) and erroneously included in the records-direct CRF (Table S2), which were then removed from scoring.

After escalation and correction, 3292 variables were available for comparison. As determined by the platform CRF, the Platform had an overall precision of 97.58% (2619/2684) (Figure 2) when compared to manual chart review. When stratified by variable type, precision levels ranged from 91.98% (tumor status) to 100% (histologic type, grade, lab result, and most recent performance status).

The overall recall capability of the platform was 81.48% (2618/3213) (Figure 2). By individual variable, recall ranged from 58.15% (adverse events) to 95.00% (breast cancer medication) (Table S2).

Additional performance measures

The number of variables extracted by the platform varied by clinical variable type (Table S2) and reflects clinical expectations. Primary Diagnosis, which is captured as a single variable regardless of the duration, had a relatively low occurrence (mean of 2.2 per individual). A mean over 1 is expected as this variable type includes not just the initial breast cancer diagnosis, but periods of NED (No Evidence of Disease) and any recurrences. Conversely, variable types that are expected to occur multiple times during the patient journey had higher average numbers of variables per individual, for example anti-cancer medications (mean of 4.4) and tumor statuses (mean of 5.2). Biomarker data had the greatest number of variables, with an average of 12.7 instances of biomarker data available per patient, which reflects the presence of multiple tests and/or broad genomic panels with multiple results.

Table 1. Validation cohort.

Characteristic	*n* = 50
Age at diagnosis, Mean years (SD; range)	47.7 (9.3; 29-69)
Histology, no. (%)	
Invasive ductal carcinoma	31 (62%)
Invasive lobular carcinoma	5 (10%)
Invasive ductal carcinoma and Invasive lobular carcinoma	4 (8%)
Inflammatory	2 (4%)
Metaplastic	2 (4%)
Other/unknown	6 (12%)
Molecular type, no. (%)	
HR + /HER2 +	3 (6%)
HR + /HER2 −	34 (68%)
HR − /HER2 +	4 (8%)
HR − /HER2 −	9 (18%)
Unknown	0 (0%)
Germline *BRCA* type, no. (%)	
BRCA negative	32 (64%)
BRCA positive	1 (2%)
Unknown	17 (34%)

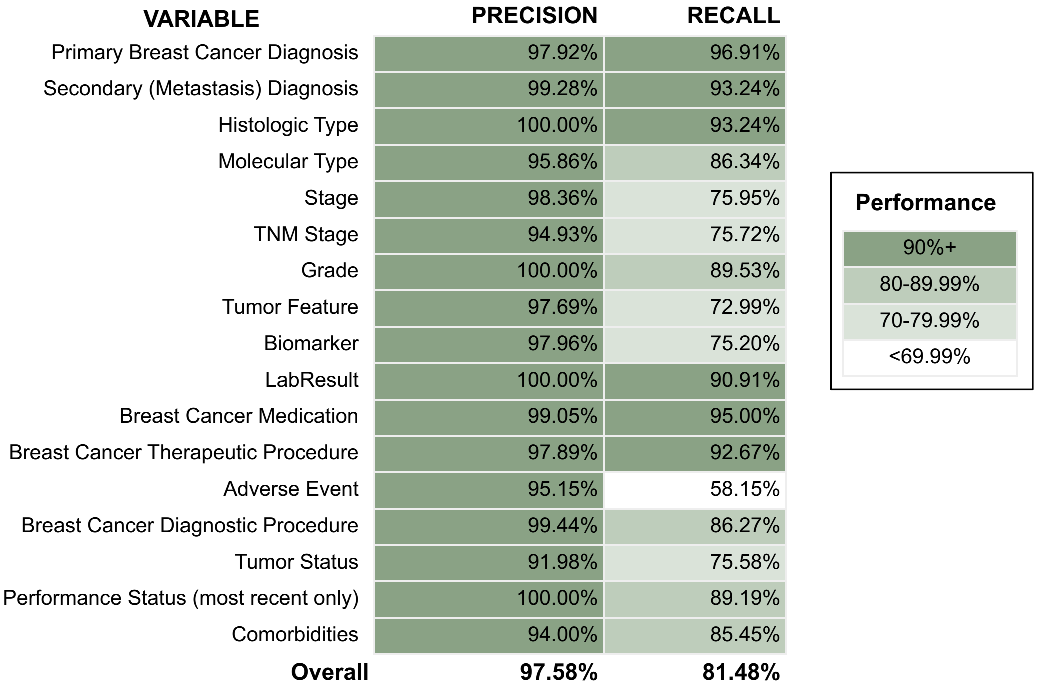
Demonstration of clinical discovery

Separately, we explored the ability for platform-extracted data to reveal clinical insights. Within a cohort of 1011 patients with breast cancer with a similar disease profile to those in the SEER database (Table 2), we identified 194 patients who developed distant metastases in the bone, brain, liver, or lung and had clinical data available prior to their metastatic breast cancer diagnosis. In this discovery cohort, the average age at initial diagnosis was 43 (SD: 9.3; Range: 29-69) and most patients (73.9%, 144/194) had invasive ductal carcinoma (Table 2). Patients most frequently had the HR + /HER2 − molecular type (62.1%; 121/194). A small percent (7.7%; 15 out of 193 individuals with documented BRCA results) of patients had a pathogenic germline variant in either *BRCA*1 or *BRCA*2 documented in their medical records. Metastases were most common in bone tissue (Table S3). Metastatic tumors in bone and brain were more common in patients with HR + molecular type compared to HR − molecular type (Table S4).

The average (range; SD) TDM was 2295.1 (31-11 985; 2020.9) days across the discovery cohort. Patients with HR − /HER2 − molecular type had the shortest average (SD) TDM (1187.4 [1129.4] days), significantly shorter (*P* <.0001) than those with the HR + /HER2 − molecular type which had the longest (2242.5 [2042.5] days) (Figure 3A). When the histologic type of the tumor was considered, patients with ductal carcinomas had an average (SD) TDM of 2270.1 (1796.6) days, and those with lobular carcinomas had an average (SD) TDM of 2812.0 (2318.5) days (Figure 3B). Patients with the rarer metaplastic and inflammatory histologies had an average TDM of 739 (354.8) days and 1005.8 days (166.2), respectively (Figure 3B), and these were both significantly shorter than the average TDMs for lobular and ductal types (*P* <.005).

By *BRCA* germline status, individuals with pathogenic germline variants in genes *BRCA*1 or *BRCA*2 (or reported as "*BRCA* positive" in their medical records) had an average (SD) TDM of 1043 (400.8) days, significantly shorter (*P* =.0007) than individuals testing negative for *BRCA*1/2 variants, who had an average (SD) TDM of 2255.9 (1822.6) days (Figure 3C). By site of distant metastasis, a liver metastasis was associated with the shortest average (SD) TDM at 2131.4 (1974.5) days and a brain metastasis was associated with the longest average (SD) TDM at 2693.8 (2407.6) days (Figure 3D).

In multivariate analyses of histology and molecular type, the shortest average TDM was observed among patients with combined ductal and lobular histologies and HR − /HER2 +



VARIABLE	PRECISION	RECALL
Primary Breast Cancer Diagnosis	97.92%	96.91%
Secondary (Metastasis) Diagnosis	99.28%	93.24%
Histologic Type	100.00%	93.24%
Molecular Type	95.86%	86.34%
Stage	98.36%	75.95%
TNM Stage	94.93%	75.72%
Grade	100.00%	89.53%
Tumor Feature	97.69%	72.99%
Biomarker	97.96%	75.20%
LabResult	100.00%	90.91%
Breast Cancer Medication	99.05%	95.00%
Breast Cancer Therapeutic Procedure	97.89%	92.67%
Adverse Event	95.15%	58.15%
Breast Cancer Diagnostic Procedure	99.44%	86.27%
Tumor Status	91.98%	75.58%
Performance Status (most recent only)	100.00%	89.19%
Comorbidities	94.00%	85.45%
Overall	**97.58%**	**81.48%**

Figure 2. Performance of medical record extraction platform. Performance of the data-extraction platform was determined by comparing case report forms completed using platform-extracted data or medical record manual review. Variables identified in both the records-direct CRF and the platform CRF were designated as true positives (TPs), variables identified in the records-direct CRF only were considered false negatives (FNs), and variables identified in the platform CRF only were considered false positives (FP). Precision was calculated as the number of TP divided by the sum of TP and FP and recall was calculated as the number of TP divided by the sum of TP and FN.

Table 2. Characteristics of metastatic breast cancer patients in the discovery cohort and the SEER database.

Characteristic	Breast cancer patients in SEER[a] database	Breast cancer patients in platform database ($n = 1011$)	Metastatic breast cancer patients in platform database ($n = 194$)
Age at diagnosis, mean (range)	NA	47 (23-80)	43 (26-73)
Histology, no. (%)			
Invasive ductal carcinoma	319 963 (72.4)	745 (74.3)	144 (73.9)
Invasive lobular carcinoma	43 930 (9.9)	80 (8.0)	17 (8.7)
Invasive ductal carcinoma and Invasive lobular carcinoma	43 090 (9.7)	32 (3.2)	16 (8.2)
Metaplastic	NA	36 (3.6)	6 (4.6)
Papillary	3326 (0.8)	4 (0.4)	0 (0)
Other or unknown	NA	114 (11.3)	9 (4.6)
Molecular type, no. (%)			
HR + /HER2 +	NA (10.0)	128 (12.7)	19 (9.7)
HR + /HER2 –	NA (68.0)	552 (54.9)	121 (62.1)
HR – /HER2 +	NA (4.0)	59 (6.0)	10 (5.1)
HR – /HER2 –	NA (10.0)	152 (15.1)	19 (9.74)
Unknown	NA (7.0)	115 (11.3)	26 (13.3)

[a] Surveillance, Epidemiology, and End Results Program data from 1975 to 2017. For some SEER data, only percentages are available.
NA, not available.

molecular type (633 days), metaplastic histologic type and HR + /HER-2 molecular type (706 days) and metaplastic histologic type and HR – /HER – (Figure 4). By histology and site of metastasis, the shortest average TDMs were among patients with metaplastic histology and metastasis to bone tissue (522.5 days), combined ductal and lobular histology and metastasis to lung tissue (573.3 days), and metaplastic histology and metastasis to lung tissue (610.7) (Table S4). By molecular type and site of metastasis, the shortest average TDMs were among patients with HR – /HER2 – molecular type and metastasis to the lung (940.0 days), liver (999.5 days), and brain (1217.0 days). Finally, patients who were *BRCA* positive had shorter average TDM for all sites of metastasis (Table S4). No multivariate results were statistically significant.

Discussion

In this validation and demonstration study, we showed that a novel EHR data-extraction platform can reliably identify clinical variables in medical records, including from unstructured data in clinician notes and pathology reports, with a

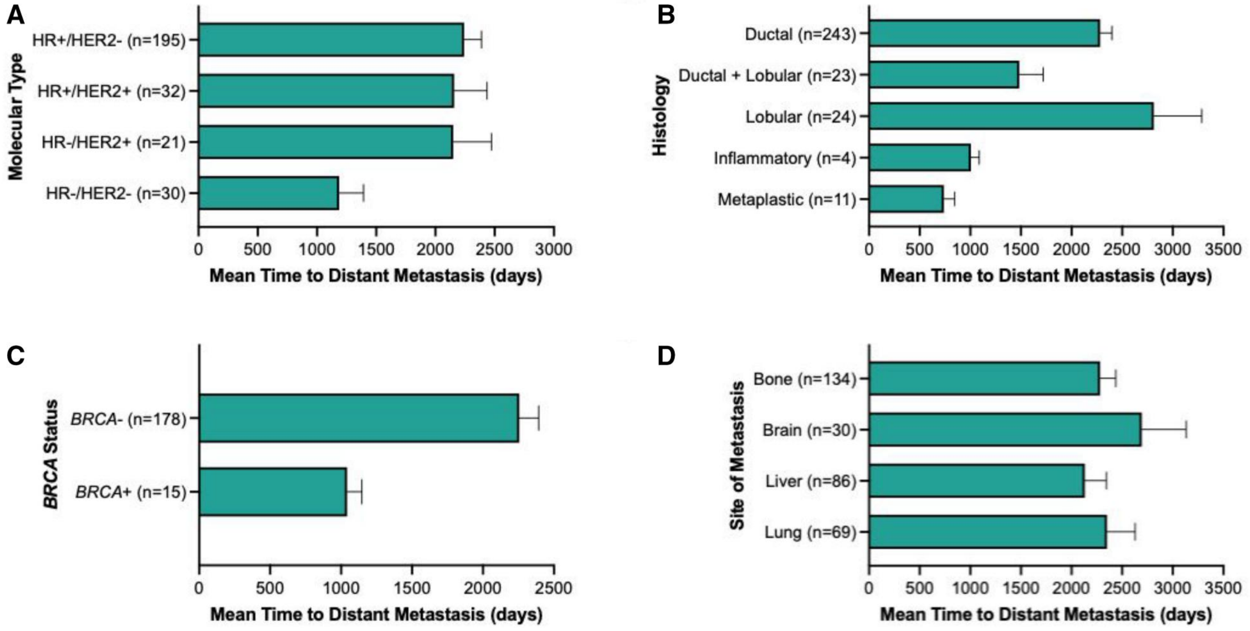


Figure 3. Tumor characteristics associated with shorter average time to distant metastasis. (A) molecular type, (B) histologic type, (C) *BRCA* status, and (D) location of metastasis. Error bars show standard error of the mean.

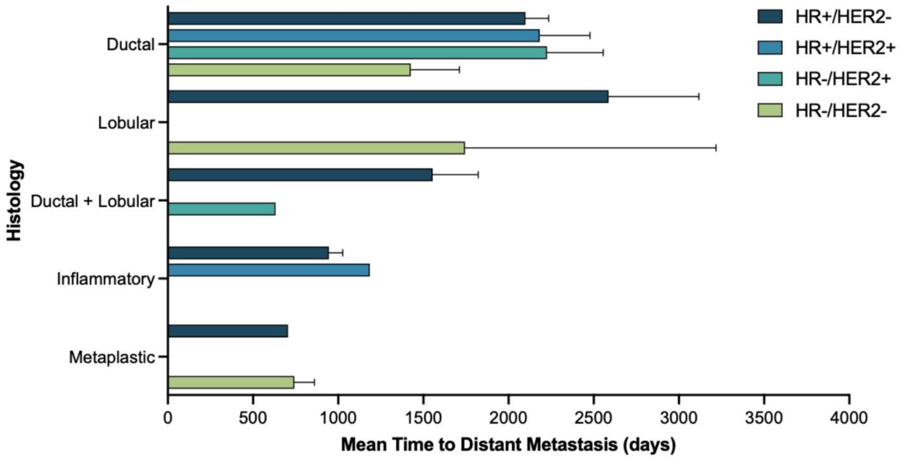


Figure 4. Multivariable analyses of tumor characteristics with time to distant metastasis. Error bars show standard error of the mean.

high level of precision (97.58%) and recall (81.48%). The platform-extracted data is also relevant to the study population of patients with metastatic breast cancer as shown through our comparison to the SEER registry. Finally, we demonstrated how this platform-extracted RWD can provide clinical insights, as the platform-extracted data revealed statistically significant associations between tumor characteristics and the average TDM for patients with metastatic breast cancer.

In addition to its utility for regulatory applications, the clinical data collected by the platform can also be used by patients to better track their diagnosis and treatment history. Patients living with cancer or rare disease face serious information burdens, multiple clinic or hospital visits, challenging terminology, and complex treatment options,[10] and a prior study has shown that nearly all patients appreciate greater access to their clinical information.[11] The patient-centered approach of the platform described here could be a solution to this issue, as it allows patients to leverage their entire medical history during their treatment journey.

RWD has long been used to assess drug safety, but now has the potential to address significant gaps in drug development.[12] New medicines and other medical products regulated by the United States government are largely tested through clinical trials in controlled settings that are expensive and cumbersome and may not fully reflect the real-world experience of individuals who may one day use those products. In addition, clinical trials are typically restricted to regions with academic medical centers where the patient population may not reflect the true diversity of individuals affected by a disease.

RWD can come from many sources, including insurance claims data, epidemiological survey data, patient reported outcomes, and aggregated EHR data. Claims data can provide longitudinal information about individuals who are continuously enrolled in specific health insurance plans, but has been shown to lack details on clinical variables, interventions, and outcomes sufficient for research use,[13] and will likely have missing data if individuals change insurance providers or programs.[7] Epidemiological survey data such as the SEER database have limited treatment information and capture only a snapshot of a given individual's health, thereby losing longitudinal information. EHRs contain detailed and longitudinal information about a patient's care, and may even include diagnosis and procedure codes found in claims data, but can have major shortcomings. For example, the clinical notes written by healthcare providers may contain a wealth of information but the unstructured format of this data type renders it nearly useless for large cohort analyses. However, others have shown that methods for computationally analyzing free text data can identify patient outcomes such as cancer metastases and other cancer events as well as adverse drug events.[2,14,15] Another potential shortcoming of EHRs is the fact that they are typically siloed to individual institutions precluding longitudinal study of patients who receive care at multiple institutions, resulting in significant gaps in a patient's treatment journey over time.

The novel Platform described in this study can resolve these shortcomings, as it draws records from all institutions identified by a patient and uses computational methods to identify and give structure to clinical variables, which are all confirmed by oncology nurse annotators. An additional advantage of the platform is the close involvement with patients, whose involvement begins with their personal directive for platform experts to request records from any institution where they have received care, and may include permission to be re-contacted for additional medical records collection or for follow-up surveys (with financial compensation) that may address chronic symptoms and quality of life. Patients also have the ability to share their data with other clinicians for second opinions and with clinical trial matching services. Although this study focused on patients with breast cancer, the clinical-extraction platform supports medical record extraction for several other solid cancer types, hematologic malignancies, and neurodevelopmental disorders and has the capability to expand and accommodate any health condition.

While the precision of all variables in this study was uniformly high (94.00%-100.00%), we observed a wide range of recall performance across variables in the validation study (58.15%-96.91%). In particular, adverse events had only 58.15% recall, which may reflect challenges for the natural language processing algorithm to identify causative statements. For example, "anemia" may be either a comorbidity or an adverse event, depending on how it is described in the medical record. While inter-rater reliability was not assessed as part of these analyses, this measure is relevant for data quality and it will be reported for the platform dataset in a future study.

In addition to its potential utility as a source of RWE for regulatory filings, this medical record-extraction Platform can reveal clinical insights as demonstrated by our analyses of TDM in breast cancer patients. While it has been previously demonstrated that HR+/HER2− breast cancer is less likely to spread to the brain than the other subtypes,[16] and that metaplastic breast cancer is more likely to be HR−/HER2−,[17] our analyses also revealed some novel findings that could inform more personalized screening recommendations for patients with breast cancer. In particular, individuals with the shortest average TDMs included those with metaplastic and inflammatory subtypes, HR−/HER− tumor molecular types, and germline variants in *BRCA1* or *BRCA2*. Our findings suggest that histology, molecular type, and genetic risk factors may be worth considering when selecting the frequency and timing of screening. Future multivariate analyses with larger n-values will continue to explore potential relationships between these clinical factors.

There are limitations to this study worthy of consideration. Given that our data set is based on real-world clinical experience of individuals and accommodates more than 1 variable per type (for example, any one individual may have zero to numerous medications), it was not possible to establish true negatives in our performance analyses and therefore we could not assess specificity. While it is not possible to generate true negatives at the variable level, future analyses will characterize specificity at the variable type level (for example, if a given case has no medications it can be considered a true negative for that variable type).

In addition, our validation was conducted with a small sample size ($n = 50$) due to the intensive personnel requirements for conducting manual chart reviews. Finally, our validation study does not account for source text quality as clinical variables may be sourced from any type of document (see Figure 1), and medical records (and the structured data extracted from them) may not contain all the relevant medical and patient information. In addition, the study was focused on breast cancer patients, and while the platform is designed to flexibly accommodate other indications, future studies will be needed to quantify the precision, recall, and clinical utility of these data more broadly.

Conclusion

A novel EHR data-extraction platform can produce structured datasets with high precision and recall and the resulting data can be used as RWD in regulatory filings or clinical discovery. Given that the platform uses disease-specific models, future efforts should validate the platform in other oncology and non-oncology patient groups.

Acknowledgments

The authors thank Robert L. Nussbaum, Deven McGraw, Flavia Facio, and Luc Cary for their thoughtful review and comments on this manuscript. In addition, we thank the patients who contributed their data for research.

Author contributions

This study was performed under the supervision of authors Alexandra Berk, Amanda Nottke, and Lisandra West-Odell, who provided significant contributions toward its design and implementation. Authors Amanda Nottke, Sophia Alan, Lura Henderson, Heather Sage, and Jessica Taylor conducted the manual medical records reviews and completed the case report forms. Authors Amanda Nottke, Anthony B. Cardillo, Hana E. Littleford, Lisandra West-Odell did the discovery cohort analyses. Amanda Nottke and Susan Rojahn drafted the manuscript, Amanda Nottke, Susan Rojahn, and Elise

Brimble generated figures and tables, and all authors provided critical review of the manuscript.

Funding

The initial design and execution of the validation section of this study was supported by Pfizer, Inc.

Conflicts of interest

All authors are current or former employees and shareholders of Ciitizen and/or Invitae, which owns the medical record data extraction platform.

Data availability

The data underlying this article cannot be shared publicly due to the privacy of individuals that participated in the study. De-identified data will be shared on reasonable request to the corresponding author.

References

1. Corrigan-Curay J, Sacks L, Woodcock J. Real-world evidence and real-world data for evaluating drug safety and effectiveness. *JAMA*. 2018;320(9):867-868.
2. Jensen K, Soguero-Ruiz C, Oyvind Mikalsen K, et al. Analysis of free text in electronic health records for identification of cancer patient trajectories. *Sci Rep*. 2017;7:46226.
3. Jiang G, Dhruva SS, Chen J, et al. Feasibility of capturing real-world data from health information technology systems at multiple centers to assess cardiac ablation device outcomes: a fit-for-purpose. *J Am Med Inform Assoc*. 2021;28(10):2241-2250. https://academic.oup.com/jamia/article-abstract/28/10/2241/6328966
4. Abernethy AP, Gippetti J, Parulkar R, et al. Use of electronic health record data for quality reporting. *J Oncol Pract*. 2017;13(8):530-534.
5. Food and Drug Administration, U.S. Department of Health and Human Services. Real-world data: assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products; Draft guidance for industry. Published Online First: September 2021. Accessed: February 12, 2023. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory
6. Center for Drug Evaluation, Research. Considerations for the use of real-world data and real-world evidence to support regulatory decision-making for drug and biological products. U.S. Food and Drug Administration. Accessed September 26, 2022. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-real-world-data-and-real-world-evidence-support-regulatory-decision-making-drug
7. Penberthy LT, Rivera DR, Lund JL, et al. An overview of real-world data sources for oncology and considerations for research. *CA Cancer J Clin*. 2022;72(3):287-300.
8. Praxis Precision Medicines. 2022. Invitae's Real-world Ciitizen Data Utilized in Praxis Precision Medicines' PRAX-222 IND Filing [Press release]. https://investors.praxismedicines.com/news-releases/news-release-details/invitaes-real-world-ciitizen-data-utilized-praxis-precision
9. Kern DM, Barron JJ, Wu B, et al. A validation of clinical data captured from a novel cancer care quality program directly integrated with administrative claims data. *Pragmat Obs Res*. 2017;8:149-155.
10. Holdsworth LM, Zionts D, Asch SM, et al. "Along for the ride": a qualitative study exploring patient and caregiver perceptions of decision making in cancer care. *MDM Policy Pract*. 2020;5(1):2381468320933576.
11. Salmi L, Dong ZJ, Yuh B, et al. Open notes in oncology: patient versus oncology clinician views. *Cancer Cell*. 2020;38(6):767-768.
12. Zhang J, Symons J, Agapow P, et al. Best practices in the real-world data life cycle. *PLoS Digit Health*. 2022;1(1):e0000003.
13. Neugebauer S, Griesinger F, Dippel S, et al. Use of algorithms for identifying patients in a german claims database: learnings from a lung cancer case. *BMC Health Serv Res*. 2022;22(1):834.
14. Wasylewicz A, van de Burgt B, Weterings A, et al. Identifying adverse drug reactions from free-text electronic hospital health record notes. *Br J Clin Pharmacol*. 2022;88(3):1235-1245.
15. Young IJB, Luz S, Lone N. A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. *Int J Med Inform*. 2019;132:103971.
16. Brosnan EM, Anders CK. Understanding patterns of brain metastasis in breast cancer and designing rational therapeutic strategies. *Ann Transl Med*. 2018;6(9):163.
17. Hu J, Zhang H, Dong F, et al. Metaplastic breast cancer: treatment and prognosis by molecular subtype. *Transl Oncol*. 2021;14(5):101054.

JAMIA Open, 2024, 7, 1–8
https://doi.org/10.1093/jamiaopen/ooae041
Research and Applications